ESC-TR-2004-084

Final Report
Pilot Corpus for Multisensor Speech Processing

October 2003

Prepared by
J.D. Tardelli – Director, Digital Speech Processing
ARCON Corporation

For MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420-9108

Under Purchase Order BX8547

Issued 6 January 2005

This work was sponsored by the Defense Advanced Research Projects Agency, Advanced Technology Office, under Air Force Contract F19628-00-C-0002.

This report is based on studies performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. This work was sponsored by the Defense Advanced Research Projects Agency, Advanced Technology Office, under Air Force Contract F19628-00-C-0002.

The ESC Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Gary Tutungian
Administrative Contracting Officer
Plans and Programs Directorate
Contracted Support Management

Abstract

The estimation of speech parameters and the intelligibility of speech transmitted through low-rate coders are severely degraded when there are high levels of acoustic noise in the speaking environment. The application of nonacoustic and nontraditional sensors, which are less sensitive to acoustic noise than the standard microphone, is being investigated to address this problem, under the DARPA Advanced Speech Encoding program. Sensors investigated include the General Electromagnetic Motion Sensor (GEMS) and the Physiological Microphone (P-mic). In order to enable this research, a Pilot Corpus with simultaneous recordings from multiple sensors has been collected by ARCON Corporation, under subcontract to MIT Lincoln Laboratory. This report describes the corpus collection, including: corpus structure, acoustic noise environments, speech materials, the sensors, and baseline intelligibility evaluations. The corpus includes Diagnostic Rhyme Test (DRT) word lists, sentence lists, and Consonant Vowel Consonant (CVC) nonsense words. Noise environments include: M2 Bradley Fighting Vehicle (M2), Military Operations in Urban Terrain (MOUT), UH- 60 Blackhawk Helicopter (BH), and a Military Command Enclosure (MCE). This pilot corpus has been utilized by a number of DARPA-sponsored research teams for R&D on advanced speech encoding exploiting multiple sensors in the military noise environments.

Table of Contents

List of Figures and Tables

1. Overview

The Advanced Speech Encoder (ASE) Pilot Corpus produced by ARCON Corporation for MIT Lincoln Laboratory consists of scripted and conversational speech produced by 10 male and 10 female talkers in a wide variety of acoustic noise environments. This speech was recorded synchronously for an array of transducers that consisted of both acoustic and nonacoustic sensors. The scripted material consisted of vowel, word, and sentence material. The conversational material was generated using an interlocutor and situational scenarios. A military scenario was also recorded in a subset of the acoustic noise environments. The transducers used included an acoustic calibration microphone, an acoustic resident microphone, an electroglottograph (EGG), two "physiological mics, 'P-mics,'" and a two-channel General Electromagnetic Movement Sensor (GEMS).

The project consisted of two phases. During Phase 1, the sensor limitations were studied and procedures were developed for Phase 2, Pilot Corpus Recording. The use of human subjects in high acoustic noise environments with experimental transducers required ARCON to submit to the regulations of an Independent Review Board (IRB). In fact, two IRBs were involved, the New England Independent Review Board (NEIRB) and the MIT Committee on the Use of Humans as Experimental Subjects (COUHES). The final Pilot Corpus contained data from 20 subjects. In addition there were two subjects that failed to show up for scheduled sessions and three subjects that dropped out of the program for various reasons. There were no adverse incidences in the program.

Procedures developed during Phase 1 were used for a small group of initial talkers and a *Preliminary Corpus* was generated and distributed. Several problems were found in the *Preliminary Corpus* and modifications were made to the process to address these problems.

The final product of this effort consisted of

- Pilot Corpus
- Talker Demographic Database
- Base-Line Intelligibility Test Results
- Post-Processed Equivalent Peak-Level Tables
- Record Logs
- Demonstration Materials – Military Scenario Recording

2. Preliminary Corpus and Modifications

The problems found with the *Preliminary Corpus* were

- High GEMS background noise
- Low GEMS SNR
- Incompleteness
- Byte order inconsistency
- Signal spikes
- Hum and hiss
- Low gain on some resident mics

The incompleteness and inconsistencies of the data format were associated with the compressed time frame and rush to provide the ASE participants with data. These problems were eliminated in the final Pilot Corpus. Many of the GEMS signal problems and signal spikes were traced to bad cable and loose connector cases.

The low gain of some resident microphones was traced to preamplifier deficiencies. The problems were addressed by

- Modified recording procedures
- Introduction of extra level-adjustment checks
- Use of an extended GEMS tuning period
- Audio monitoring of GEMS and P-mics
- Cable and connection checks and rechecks
- Changes to resident microphone interfaces

3. Pilot Corpus Structure

The Pilot Corpus is provided with the following data format:

- Raw headerless audio files
- Mono
- 16 bit
- 0.48k extension = 48kHz sample rate
- 0.16k extension = 16kHz sample rate
- Intel byte order (LSB, MSB)

Data for 20 talkers (M0, M1, … M9 and F0, F1 … F9) are contained in their separate file folders. The total data for each talker takes approximately 4GB.

Data from nine acoustic noise environments is provided for each talker. The following are the environments and their file folder names:

Quiet – E0_qt
Office – E1_off
USAF Mobile Command Shelter (MCE) – E2_mce
M2 Bradley Fighting Vehicle (high noise level) – E3_m2h
M2 Bradley Fighting Vehicle (attenuated 40dBC) – E4_m2l
Military Operations on Urbanized Terrain MOUT (high noise level) – E5_moh
Military Operations on Urbanized Terrain MOUT (attenuated 40dBC) – E6_mol
UH-60 Blackhawk Helicopter in flight (high noise level) – E7_bhh
UH-60 Blackhawk Helicopter at idle (attenuated 40dB from high) – E8_bhl

Processed files are contained in the /final folder for each talker. These files contain

- One four-page Diagnostic Rhyme Test
- Two ten-sentence Harvard PB lists
- One twenty-word Consonant-Vowel-Consonant list in a carrier phrase
- One ten-word "sustained" vowel list
- One five-minute conversational sample

These files are provided for all talkers and environments except for Quiet. The Quiet environment files were recorded for DRT and CVC material only.

Supplemental material is contained in the /supp folder for each talker. This material has not been fully post-processed and consists of the "outtakes" of the /final material. The supplementary material may not match the scripted material and can contain incomplete and/or noisy data. The supplementary DRT material is for single pages only. No editing except for the removal of digital zero sections has been performed on this material.

Each file set contains the data outputs of from seven to eight sensors. These sensors are to some extent environment dependent. Sensors and their files' sample rate are:

Sensor 0, 48K – the environment's primary resident sensor
Sensor 1, 48K – the B&K reference microphone
Sensor 2, 16K – the Electroglottograph
Sensor 3, 16K – the P-mic at the throat location
Sensor 4, 16K – GEMS Channel A
Sensor 5, 16K – GEMS Channel B
Sensor 6, 16K – the P-mic at the forehead location
Sensor 7, 48K – the environment's secondary resident microphone when available.
Sensor 8, 48K – Sennheiser MHD224 audio of the second party for conversations
Sensor 9, 48K – the B&K reference microphone for test tones

All files use the following naming convention: GnEMMMMS.rate, where

G = gender, either M(ale) or F(emale)
n = talker number 0 to 9
E = environment 0 to 8 as listed above
MMMM = Material:
VWLn for vowel lists 1 to 3
HSnn for Harvard Sentence lists 01 to 72
DRT list number, ie 301A
CVnn for Consonant-Vowel-Consonant lists 01 to ??
CNVn for conversations 1 to 8
TONn for calibration test tones
S = sensor 0 to 9 as listed above
rate = sample rate extension, either 48K or 16K

Each audio file (0.48k) has been equalized using the Equivalent Peak Level method. For each audio file there is a ".epl" text file that contains the equalization level applied to the file along with the speech-to-noise level for that file as measured by ARCON's EPL process. The algorithm's accuracy in measuring SNR declines below 5 dB. These values should be used with caution. Values of SNR lower than –6 dB indicate a failure of the algorithm to measure the SNR. The calibration test tone files, GnETONnS.48k are provided for each talker-environment at the environment level in the file structure. For a limited number of cases there may be more than one file. This file is a recording of a B&K calibration test tone (Sound Level Calibrator Type 4230 – 1000 Hz at 94 dB) through the B&K calibration microphone as adjusted for the particular environment.

The "final" folders have been reviewed for completeness and content. Some of the problems seen in the preliminary corpus release have been corrected, while some still exist. Those problems associated with noisy signals and/or low signal-to-noise ratios for the talkers in the preliminary release, especially F0, still exist.

Additional information is provided as follows:

- A document detailing the Pilot Corpus content
- All talker demographic and sensor information
- GEMS placement photos for all talkers
- Lists of all scripted material used in this database
- DRT word list decode information
- Microphone and sensor information
- EPL Reference list

- Copies of the Pilot Corpus Phase 1 and 2 design-review presentations
- A short "Cheat Sheet" with environment and sensor codes
- A list of remaining file problems and exceptions
- DRT results for all "final" DRTs both NULL and MELPe processed
- CVC test results for a subset of the "final" conditions
- Military Scenario Demonstration files

4. Acoustic Noise Environments

The acoustic noise environments utilized in the ASE Pilot Corpus can be separated into three *Baseline* environments and three *40dB Difference* environments. The baseline environments provide the capability to trace back to other corpora and benign or moderate noise fields for baseline processing and comparisons. A goal of Phase 1 of the ASE Program was to demonstrate a *40 dB improvement in the intelligibility of coded speech in military noise environments.* Extensive effort went into the interpretation of this requirement and the consideration of metrics that could be used in the measurement of the requirement. This resulted in the *40dB Difference* environments. For these environments there are two states, an operational severe acoustic noise state and a benign state that is 40 dB lower in Sound Pressure Level (SPL) from the severe state.

4.1. Baseline Environments

Quiet – The ambient state of the recording room is used for the quiet environment. The SPLs within the acoustic isolation room used for all recordings measure 20 dBA and 44 dBC.

Office – A large, partitioned modern office environment. The major acoustic characteristics include background conversation, ventilation, ringing phones, and computer and printer activity. Typical SPL values are 65 dBLin , 55 dBC and 45 dBA. Ventilation noise and computer fan noise are constant throughout the noise samples. Background conversations are quite common but vary in amplitude. Keyboard activity, ringing phones, and printer activity are sparse and of short duration, although phone ringing is a dominant characteristic during its existence.

Mobile Command Enclosure (MCE) – The occupied space of an MCE is approximately 12 feet by 12 feet and contains four workstations. Near-constant background conversation and constant high-volume ventilation/cooling noise dominate the acoustic characteristics.

4.2. 40dB Difference Environments

U.S. Army M2 Bradley Fighting Vehicle - The major acoustic characteristics of the M2 include the engine noise and the tracks as the vehicle maneuvers. The vehicle was driven on a course at a consistent speed of 40mph creating SPLs of 109-112 dBA , 124-126 dBLin. The engine noise is severe and constant. Track noise variations that occur as the vehicle turns and proceeds over terrain are easily perceived. The higher frequency range they occupy separates them from the engine noise, but the impact on SPL is insignificant.

Military Operations in Urban Terrain (MOUT) - The major acoustic characteristics of the MOUT environment include small arms, grenade and mortar fire. A 55dBC SPL reflects a typical noise floor during a period of inaction and 95dB SPL is typical for moments of gunfire. Periods of action and inaction vary widely across the sample. Intense moments of automatic weapon, grenade, and mortar fire lasting between 10-20 seconds up to several minutes will be interspersed with sections of no action whatsoever or occasional shouts and footfalls. Within periods of action, intensity also varies quite widely from a single small arms shot to full squad vs. squad battle.

Black Hawk Helicopter UH-60 - There will be two conditions for this; cruise and idle (-40dB from cruise). The cruise environment represents the Blackhawk in flight at a standard cruise speed. The recording took place at the co-pilot position. The dominant acoustic characteristic is the jet engine noise. Some rotor wash is occasionally heard. The idle environment was recorded with the Blackhawk on the ground. The engines are on, but the rotors are not spinning. The helicopter is also attached to its ground-based generator, which is running. This is the standard pre-flight configuration.

5. Speech Materials

The speech materials contained in the Pilot Corpus consist of both scripted speech and conversational speech. The scripted materials consist of vowel sounds, words, and sentences. The vowel sounds are contained in lists of words with medial vowels. When the word is uttered by the talker, the talker is asked to extend the vowel sound. The word list is structured such that it contains a sample of the common vowels of North American English. The word lists used in the Pilot Corpus are both Diagnostic Rhyme Test word lists and Consonant-Vowel-Consonant word lists. This provides for intelligibility test input material, The CVC words are presented by the talker within the carrier phrase "mark the word *cvc* now." The sentence material used in the Pilot Corpus is from Harvard Sentence Lists. This provides for quality test input material. Conversational speech is generated through the use of conversational scenarios and a second talker acting as an interlocutor. These conversational scenarios were designed to provide five minutes of conversation with the focus on weighting the conversation heavily on the primary talker. Additional conversational material is contained in the Pilot Corpus in the form of the demonstration military scenarios. Unfortunately this material exists for only two talkers. These talkers are not a subset of the 20 Pilot Corpus talkers.

6. Sensors

6.1. Resident Microphones

Quiet – Altec 659A Dynamic microphone. This is a professional quality microphone designed for close micing techniques. It has been used as a baseline resident microphone for quiet-speech source recordings extensively within the DoD.

Office – STUIII Telephone handset. An electret condenser microphone housed within a typical telephone handset. This particular model was part of the STUIII systems.

MCE – M-87 noise-canceling, dynamic microphone, boom mounted to circumaural headphones that provide >13 dB passive-noise attenuation at the user's ears.

M2 – A subset of talkers used the CVC Model DH-132A, an older, passive-noise-reduction helmet equipped with the M-138 dynamic noise canceling boom microphone for a portion of their recordings. It provided >14 dB passive noise protection at the talker's ears. All other recordings used a CVC Helmet with Active Noise Reduction (ANR). Bose Product Improved Combat Vehicle Crewman (PICVC) Helmet, H-374(V)5/VRC; Second generation ANR helmet with M175 Electret noise-canceling boom microphone assembly. Specific information concerning which microphone was used for a given talker/list was included within the documentation supplied with the corpus.

MOUT – MICH Helmet with Communications Subsystem: The communication subsystem is intended to provide aural protection as well as a dual-channel communications capability. The subsystem provides aural protection, occluding and non-occluding communications, omnidirectional hearing, ear-specific communications (dual channel), low-profile microphone(s), microphone adapter for mask microphone, multiple radio and intercom adapters, and push-to-talk

access. The headset may be worn alone or with the ballistic helmet retention system and pad suspension system.

The acoustic microphone of a Telex Stinger tactical headset was simultaneously recorded to supplement the bone conduction microphone provided with the MICH helmet. This is a noise-canceling electret microphone mounted on a lightweight single-ear headset designed for Spec Ops.

Black Hawk – Acusticom 59369 dynamic noise-canceling microphone boom mounted to an SPH-4 helicopter helmet providing >13dB passive-noise attenuation at the user's ears.

6.2. Calibration Microphone

B&K – Type 4155. A prepolarized full-bandwidth omnidirectional condenser microphone designed to provide consistently flat response across a wide range of SPLs.

6.3. Electroglottograph (EGG)

The EGG used is the Glottal Enterprises Model EG2 with 35mm electrodes. This is a noninvasive device that detects changes in impedance across the subject's vocal folds during a speech vibratory cycle. The EGG employs a pair of electrodes that are placed in contact with the subject's skin on both sides of the larynx and held in place with a collar. The EGG has an interelectrode voltage of one V rms and a current of 10mA at 2 MHz. This device is a standard clinical tool used by Speech Pathologists and Linguistic Researchers. It poses no risk to the subject.

6.4. Physiological Microphone

The P-mic is optimized for hands-free use. The microphone is designed to eliminate most background noise. The P-Mic is worn like a collar and has a silicon contact sensor applied directly to the skin. A battery-operated preamplifier is included and provides a line-level signal.

6.5. GEMS

The GEMS sensor is manufactured by Aliph. It is a low-power, battery-powered, radio-frequency (RF) noninvasive device that is capable of measuring very small relative motions from the subject's tracheal wall. The GEMS sensor is a continuous-wave phase-modulated RF device that operates at a frequency of precisely 2.4 GHz. It uses a monostatic antenna configuration that utilizes reverse polarity SMA connectors. The maximum theoretical RF power output is 1 milliwatt while the effective power will depend on the efficiency of the antenna. It returns a voltage signal limited to ± 2.5 V that indicates the amplitude of motion of a periodic or quasi-periodic motion of a target in its field of view. It is capable of measuring very small (micron order) vibrations from 20 Hz to 8 kHz, both inside and outside the body. It operates on a rechargeable 7.4V Li-Ion battery and draws approximately 170 mA of power. The GEMS device is a transmitter and a receiver. In operation, it transmits less than 1 mW (0 dBm) of power in the frequency range of 2400–2500 MHz. When used as instructed by Aliph, the GEMS device is designed to operate in compliance with the RF exposure guidelines and limits set by the FCC and international health agencies. The GEMS complies with Part 15 of FCC regulations concerning the use of intentional and unintentional radiators in the 2.4GHz ISM band. The exposure due to the use of the GEMS is far below that of a cell phone. Thus, these exposures are considered within that of normal, everyday life and as such present minimal risk.

6.6. GEMS Configuration Procedures

Trials involving MIT Lincoln Laboratory and ARCON personnel were conducted during Phase 1 to understand the performance characteristics of the GEMS units and establish procedures to properly apply the sensors. Data from these trials were also submitted to Aliph personnel for guidance. The following sections describe the equipment and the procedures derived from these trials that will be implemented during talker recording.

GEMS Application/Positioning – A single GEMS 9p6+8_80 antenna is affixed to a 1-inch thick Velcro strap. EGG sensors are placed along the strap on either side of the GEMS antenna. A felt pad is installed behind the GEMS antenna so that the antenna and EGG sensor surfaces are at the same height. This strap is placed around a talker's neck so that the GEMS antenna lies flat on the skin just below the laryngeal notch. Using an oscilloscope, the GEMS output is monitored while the talker enunciates an elongated vowel sound. The GEMS antenna is repositioned by adjusting the neck strap until the best possible signal is received. The EGG unit provides a sensor indicator display. The display value will be logged and used to replicate GEMS placement throughout the multiple recording sessions.

Antenna Tuning – If a robust signal of sufficient strength is measured with the default configuration of antenna 9p6+8_80 without any cellophane tape, that will be the talker's configuration. Otherwise, a layer of cellophane tape will be applied to the surface of the antenna and the antenna repositioned on the talker. Up to 3 layers of tape will be used in an attempt to tune the antenna so that the GEMS unit outputs a robust signal. If necessary, the 9p6+8_80 antenna will be replaced by the 9p6+7_80 antenna and the tape tuning process repeated. This could be followed by a repeat of these procedures using the 9p6+6_80 antenna.

Signal Strength – Oscilloscope readings of the GEMS unit output were observed during trials conducted by ARCON Corporation. Based on these readings, a GEMS output voltage range was derived that provided a stable, undistorted signal. Signals above 3.25 Vp-p were consistently seen to have distortion in the form of clipping. Signals below 0.5 Vp-p provided little energy above the noise floor and were often unstable, changing shape and frequency. Consequently it was determined that a GEMS output operating voltage range that provides a reasonably strong signal without amplifier distortions was 1–3Vp-p.

Signal Quality – Trials involving five talkers across all antennas and up to six different tunings revealed a number of different waveforms might be output by the 2-channel GEMS for a given talker uttering the sustained vowel sound "eee" as in "beet." Digital images of these waveforms were generated and distributed among key personnel at ARCON, Lincoln Laboratory, and Aliph. From the subsequent discussions, an understanding was reached concerning what constitutes a "good" GEMS signal. Following are some examples. Figure 1 shows the 2-channel GEMS output. Figures 2 and 3 include a middle trace of a VFCA EGG output.

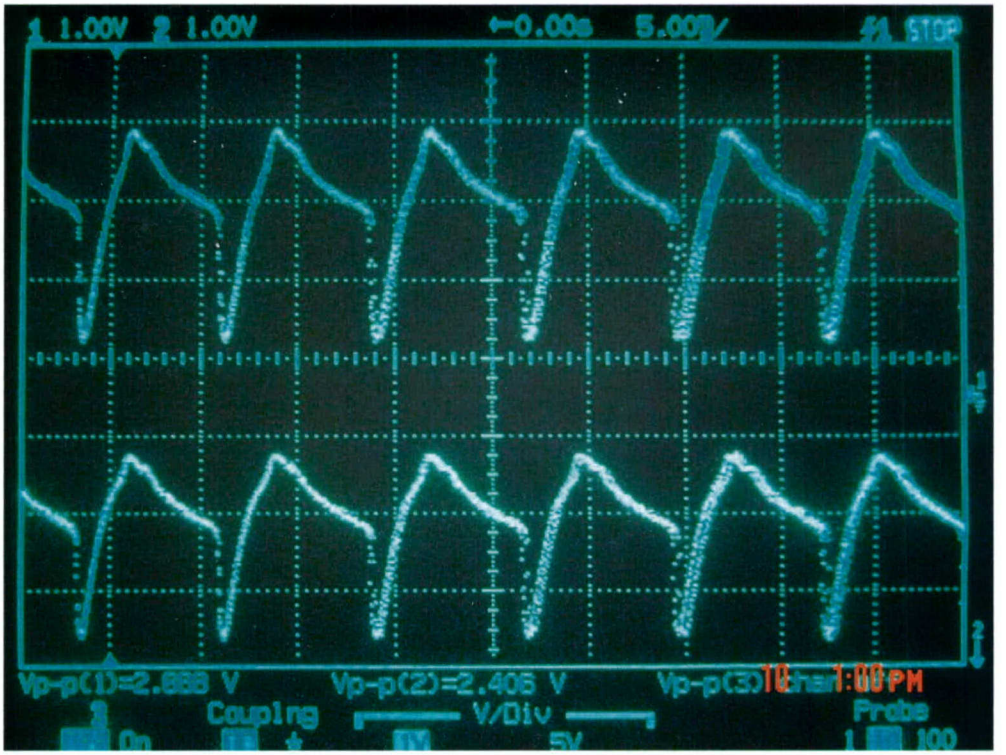

Figure 1. GEMS 2 Channel Output Sample.

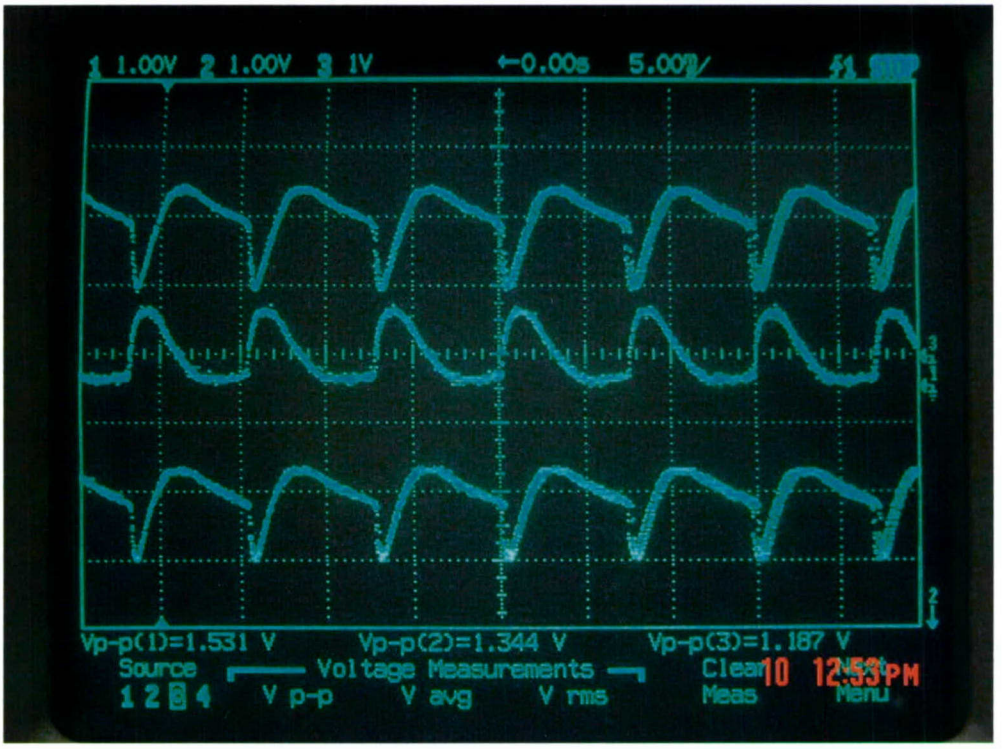

Figure 2. GEMS Sample with EGG Trace.

The previous figures display the desired response from the GEMS at the trachea. An important feature of these waveforms is they have only two zero crossings per speech period. The signals shown in Figure 3 are also good, especially the top one. These tracings look much different because the magnitude of the top signal is out of phase with the bottom, so it is upside down in comparison. However, this does not affect its qualities and it is still a "good" signal.

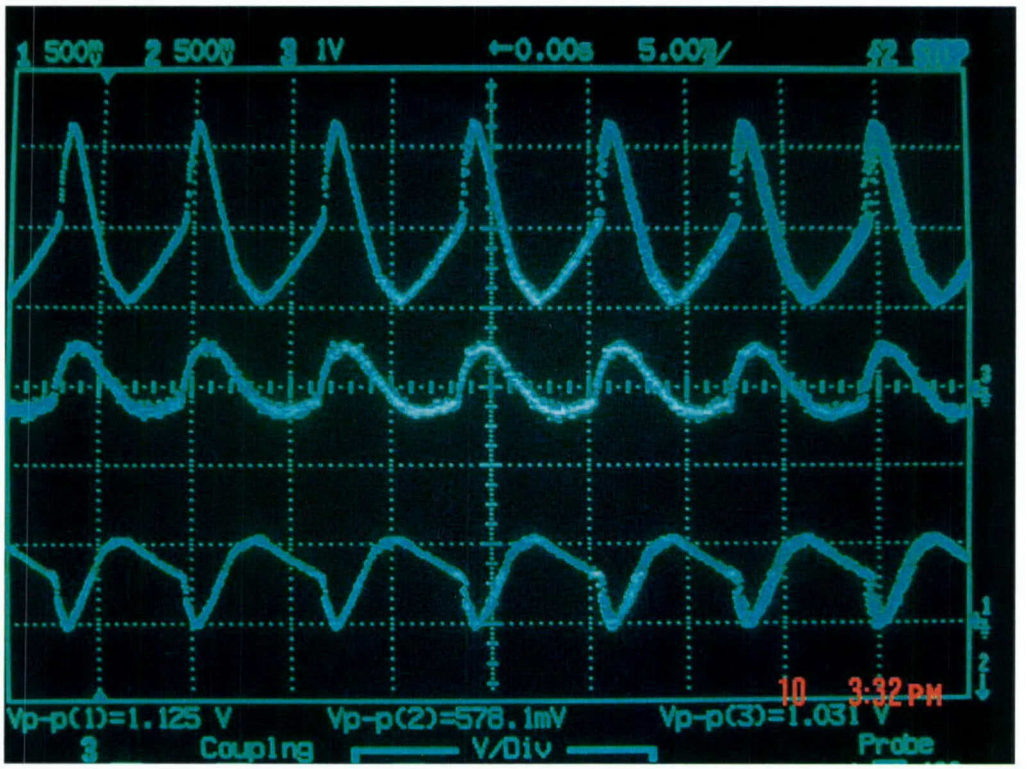

Figure 3. Inverted GEMS Sample.

7. Talkers

7.1. Human Subject Issues

The development of speech corpora in high acoustic noise environments with sensors that include active elements introduces a risk factor to the human subjects during the recording stage. This risk factor is minimized through the proper design of experimental protocols, the subject's advised consent document, the subject population, document security, and recruitment practices. An independent review of these factors is required to assure that any adverse effects to the human subject are minimized and their rights are protected. A Human Subjects Committee or Institutional Review Board (IRB) typically performs this review. For this project ARCON submitted the experiment protocol and an Informed Consent document to both the New England Independent Review Board (NEIRB) and the MIT Committee on the Use of Humans as Experimental Subjects (COUHES).

7.2. Demographic Information

The Subject Database has two portions, one that is unlikely to vary between sessions and another that does vary between sessions. The subjects provide information for this database. Some questions are optional. If the subject does not provide some of the information, the experimenter may provide their assessment of the answer. The following is a list of the information provided in the Subject Database. Items marked with a star (*) are mandatory.

Talker Portion:

Assigned ID Number (relative to program only, not SSN or driver's license) *
Sex *
Race
Age
Height
Weight
Place of birth
Native language
Years of English
Current smoker (Does the person have a "smoker's voice"?)
Dental appliance (including tongue piercing)

Session variations:

Date and time of recording *
Recent illness (temporary speech problem, e.g., congestion, head cold, allergies)
Dental-appliance changes
Dental work since last session
Comments (any unusual circumstances of session)*

8. Reproduction and Recording System

8.1. Acoustic Noise Environment Simulations

The ARCON Corporation recording facility (see Figure 4) includes an acoustic isolation room equipped with a calibrated professional audio playback system, all of which works to faithfully recreate a wide variety of acoustic noise environments within the frequency range of speech (40 – 4,000 Hz). Noise sources are stereo, full bandwidth recordings collected in the actual noise environment along with Sound Pressure Level (SPL) measurement data and reference calibration tones. These sources have been saved as 48-kHz sampled, 16-bit headerless PCM formatted files. These hard disk files are transferred digitally to outboard D/As to provide the best possible signal to noise ratio during analog conversion. The analog signal is fed through a professional audio preamplifier/splitter, a crossover, and onto power amplifiers being fed to the loudspeaker array within the isolation room. This high-power, large diaphragm loudspeaker array has been specifically configured to reproduce the immense amount of low-frequency energy that characterizes many commercial and military ground vehicles (e.g., cars, personnel carriers, tanks, etc.). The amplifier/speaker sets have been balanced to eliminate distortion due to signal saturation. Speaker arrays have been carefully positioned and calibrated to reduce the effects of acoustic phasing caused by time alignment of loudspeaker systems or interaction with room surfaces. Laboratory standard reference microphones and measuring amplifiers are used to adjust playback gains based on the SPLs and calibrated reference tones measured and recorded in the field. This procedure provides an accurate simulation of the field-recorded sound field in the acoustic chamber.

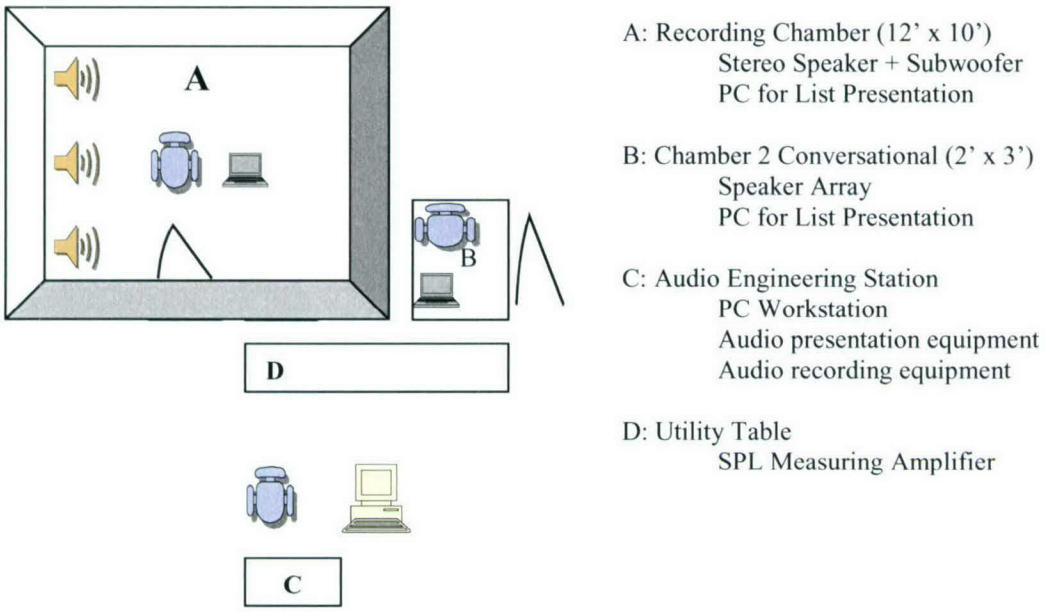

Figure 4. ARCON Corporation Speech Recording Facility.

8.2. Multitrack Recording

Signal flow within the ARCON recording facility was designed to provide all necessary functionality in a flexible configuration (see Figure 5). Professional quality matrix amplifiers were used to split and mix the resident microphone signals to provide talker sidetone, allow two-way, full duplex communication with an interlocutor, and include a method by which the recording engineer interacted with the talker in order to begin and end each specific take. The resident mic was monitored via headphones to identify talker recitation errors. Headphones and oscilloscopes were used in monitoring non-audio sensor output.

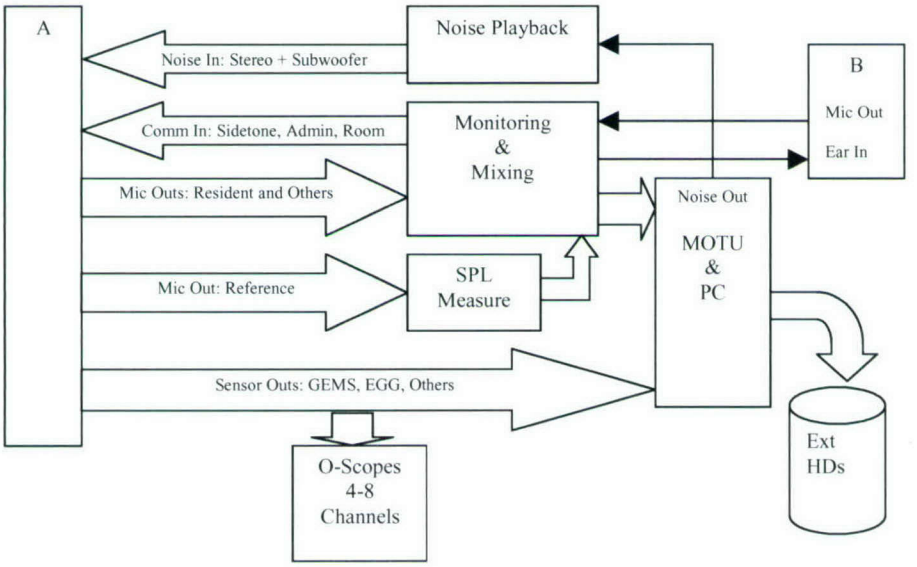

Figure 5. ARCON Corporation Speech Recording Facility ASE Database Recording Signal Flow.

Syntrillium Corporation's "Cool Edit Pro 2.0" software suite was used for noise playback and multichannel recording during talker sessions. The recording software supplied a time vs. power display for each channel. It also allowed unique sessions to be predefined for each talker/environment/configuration and unique blocks for each word or sentence list recorded. This same suite was used for editing of the raw recordings and post-processing the supplemental material.

8.3. Sensor Interfaces

GEMS, P-Mics, and the EGG sensors came equipped with preamplifiers that provided signals at or near line level, and no additional interfacing was required beyond simple cable/connector adaptors. Dynamic resident microphones also required only cabling and the use of a microphone preamplifier. Custom interfaces were built for resident microphones using an electret condenser design. Either batteries or a power supply was used to provide the correct DC voltage for each mic. Circuitry was included to properly match impedance, eliminate DC signal from the audio output, and provide a clean signal. The reference mic included its own preamplifier and battery power supply and provided a line-level signal.

9. Recording Schedule

The modified recording schedule was structured to consist of two days with two subjects being recorded on both days. The first day was allocated 50% to each subject and consisted of registration, hearing threshold testing, IRB issues, and recordings in the "baseline" environments—Quiet, Office, and MCE. The schedule for a single talker can be seen in Table 1. On the second day, two talkers are processed by alternating their recording sessions with offset break periods. This schedule can be seen in Table 2.

SCREENING DAY	
Registration	5 min
Intro to recording rooms	5 min
Intro to sensors/headsets	10min
Intro to recording schedule	5 min
Consent Agreement Intro	10 min
Hearing Screen	20min
Signed Consent	5 min
Sensor placement	20 min
Training	40 min
Quiet DRT, CVC	
Break	10 min
#1 Setup Office	
Transducer Configuration #1	5min
Scripted Recording	15min+5min
Conversational Recording	5min
Break	10min
#2 Setup MCE	5 min
Transducer Configuration #2	5min
Scripted Recording	15min+5min
Conversational Recording	5min
Break	10min
Hearing Screen	15min
Scheduling	5min
Talker 1 Screen Complete	

Table 1. Recording Schedule for Screening Day.

RECORDING DAY SCHEDULE					
Talker 1			Talker 2		
9:00	Startup/Threshold Screen	20min			
	#3 Setup M2 High	10min			
9:20	Transducer Configuration #1	10min			
9:30	Scripted Recording	15min+5min	9:30	Startup/Threshold Screen	25min
9:50	Conversational Recording	5min		Environment #3 M2 High	
9:55	Break	35min	9:55	Transducer Configuration #1	10min
			10:05	Scripted Recording	15min+5min
	#4 Setup M2 Low		10:25	Conversational Recording	5min
10:30	Transducer Configuration #2	10min	10:30	Break	35min
10:40	Scripted Recording	15min+5min			
11:00	Conversational Recording	5min		Environment #4 M2 Low	
11:05	Break	45min	11:05	Transducer Configuration #2	10min
			11:15	Scripted Recording	15min+5min
			11:35	Conversational Recording	5min
11:40	#5 Setup MOUT High	10min	11:40	Break	45min
11:50	Transducer Configuration #1	10min			
12:00	Scripted Recording	15min+5min			
12:20	Conversational Recording	5min		Environment #5 MOUT High	
12:25	Lunch Break	60min	12:25	Transducer Configuration #1	10min
			12:35	Scripted Recording	15min+5min
			12:55	Conversational Recording	5min
	#6 Setup MOUT Low		13:00	Lunch Break	60min
13:25	Transducer Configuration #2	10min			
13:35	Scripted Recording	15min+5min			
13:55	Conversational Recording	5min		Environment #6 MOUT Low	
14:00	Break	45min	14:00	Transducer Configuration #2	10min
			14:10	Scripted Recording	15min+5min
			14:30	Conversational Recording	5min
14:35	#7 Setup UH-60 High	10min	14:35	Break	45min
14:45	Transducer Configuration #1	10min			
14:55	Scripted Recording	15min+5min			
15:15	Conversational Recording	5min		Environment #7 UH-60 High	
15:20	Break	35min	15:20	Transducer Configuration #1	10min
			15:30	Scripted Recording	15min+5min
			15:50	Conversational Recording	5min
15:55	#8 Setup UH-60 Low	10min	15:55	Break	45min
16:05	Transducer Configuration #2	10min			
16:15	Scripted Recording	15min+5min			
16:35	Conversational Recording	5min		Environment #8 UH-60 Low	
16:40	Break	15min	16:40	Transducer Configuration #2	10min
			16:50	Scripted Recording	15min+5min
16:55	Hearing Screen	15min			
17:10	Break	15min	17:10	Conversational Recording	5min
			17:15	Break	15min
17:25	Hearing Screen	15min	17:30	Hearing Screen	15min
17:40			17:45	Break	15min
			18:00	Hearing Screen	15min
			18:15		

Table 2. Recording Schedule, Day 2, Two Talkers.

10. Procedures

10.1. GEMS Optimization

The antenna, tuning, amplification, and placement arrived at during a talker's GEMS configuration session (see section F) was replicated at the start of each recording session. The talker pronounced elongated vowel sounds and some sentence material. Oscilloscope monitoring during application provided feedback on signal strength and form.

10.2. Sensor Placement

Sensors applied to the talker's throat were positioned with the priority of optimizing the GEMS sensor. The EGG sensors are placed on either side of the GEMS antenna on the same 1-inch Velcro strap. This strap is attached around the talker's throat at the optimal GEMS glottal position. A P-mic is also attached at the talker's throat either above or below the GEMS/EGG sensors. The majority of talkers have the P-mic below the GEMS/EGG. Signal overload was less prevalent at this position, and physical interference with the GEMS/EGG was minimized. A small subset of talkers was recorded with the P-mic above the GEMS/EGG. A second P-mic was applied to all talkers at the center of the forehead. The reference mic was mounted to a tripod stand and placed approximately 1/2" from the talker's lips and slightly below. The resident mic was the priority acoustic sensor and was either boom- or stand-mounted and placed within ½" of the talker's lips.

10.3. Talker Position

Talkers were seated at the optimal position within the sound room for accurate noise simulation, i.e., the position where the noise-field playbacks were calibrated. Talkers faced away from the noise-playback loudspeakers and toward a desk and laptop PC that presented the word and sentence lists. Talkers were seated in order to maintain a comfortable and consistent position—feet flat on the floor, back against the back of the chair, arms on the armrests. The height and distance of the PC were adjusted to keep the talker's head level.

10.4. Recording Engineer Tasks

The recording engineer was responsible for configuring the resident microphones and noise fields prior to the talker session. At the beginning of each session, the engineer assisted in applying sensors to the talker, assuring their proper operation and adjusting analog levels into the recording system. During the session, the engineer operated the software suite for noise playback and talker recording, continuously monitoring all sensor levels and signal quality. At the end of each session, the engineer saved all tracks to disk.

10.5. Producer Tasks

The producer performed the GEMS configuration efforts and applied the various sensors to the talker, replicating the optimal GEMS placement and verifying signal strength and form. The producer was the only person in contact with the talker during a session and maintained a dialog between recordings to insure the talker was comfortable, understood the task at hand, and maintained proper positioning and sensor placement. Talkers were coached when necessary on pronunciation, positioning, and laptop operation. The producer monitored the word and sentence lists via the resident microphone signal as they were recited. The producer then determined, based on proper pronunciation, if the recording take was good or if it had to be repeated. The producer also tracked and logged elapsed session and noise-exposure times.

11. Post Processing

The post-processing consists of the following activities:

- Archival and backup
- Working copies of the multitrack raw recordings edited
- DRT pages edited for concatenation
- WAV headers removed
- Supplemental material zero stripped
- DRT pages concatenated
- EPL measures and Acoustic Gain adjustment to all audio signals
- Nonacoustic sensor data downsampled to 16 kHz
- Files renamed and placed in hierarchical directory structure
- Test Tone files and EPL output files added to file structure

The tools used for post-processing are from either the *ITU-T Software Tool Library, ITU-T Recommendation G.19,1* or are ARCON proprietary software. The ITU-T software is used for cutting files, concatenating files, and downsampling. The subroutine calls are:

astrip -q -sample -start -22 input output
concat -f page1 page2 page3 page4 Gnexxxxs.48k
filter -q -down HQ3 input.48k output.16k
filter -q -down HQ2 input.16k output.8k

The characteristics of the downsampling filters are provided in Figure 6.

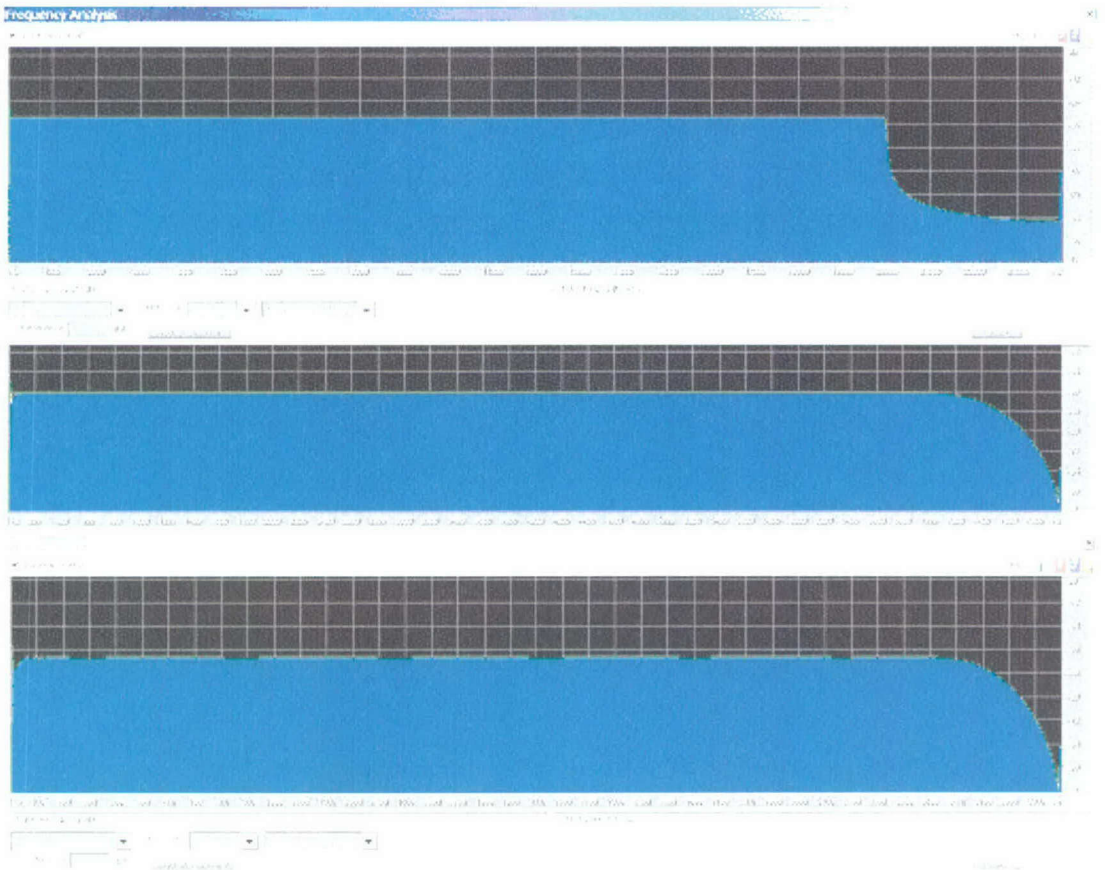

Figure 6. ITU-T "*filter HQ2 and HQ3*" Characteristics.

11.1. Equivalent Peak-Level Adjustment of Audio Channels

The gains of all acoustic microphone files were adjusted to provide consistent speech presentation levels across all talkers, microphones, and environments based on EPL measurements. An automation script was written to measure, adjust, and re-measure files until reaching a target EPL value. Speech-to-noise measures were recorded. All level adjustments are available in the /EPL folder of the Pilot Corpus.

12. Exceptions

12.1. Resident Microphone

The PICVC system in the M2 and the MICH system in the MOUT (see section F) were obtained after the recording effort began and were not available for the first six talkers. A subset of their material, the DRTs and CVCs, were rerecorded in later sessions when possible. As mentioned in section F, the Stinger mic was included in the MOUT environment as the MICH provided only a bone conduction sensor. Specific information concerning which microphone was used for any given talker/file was included within the documentation supplied with the corpus.

12.2. Sensor Problems

A number of GEMS antenna leads or connector problems occurred. These problems caused a noisy GEMS output signal. A total of four antenna cables failed during the recording effort as well as a number of instances of loose connectors. Procedures were developed to better prevent and detect these failures during the recording effort, and the corruption is mainly limited to the first six talkers recorded. Specific information concerning when a failure occurred was included within the documentation supplied with the corpus.

13. Baseline Intelligibility Evaluations

13.1. DRT

The results from DRT evaluations for each of the 20 talkers in each of the nine environments were provided for a single DRT word list in both the NULL processed and MELPe processed states. The NULL process did include the downsampling to an 8kbps, 4kHz audio bandwidth. The input to the MELPe process was the NULL process. This database of DRT results consists of the standard DRT report for 20*9*2 single-speaker systems.

13.2. CVC

The CVC test design was as follows:

- Eight conditions (Quiet, Office, BHH, BHL, M2H, M2L, MOH, MOL)
- Twelve talkers (six male plus six female)
- Two processes (NULL, MELPe)

This results in the presentation of 192 stimuli to each CVC test subject. The stimuli were divided into eight blocks of 24, with each block balanced across talkers, talker sex, processing, and condition. The blocks were randomized to minimize any possible presentation-order affect. A single CVC word in its carrier phrase was used for each condition-talker-process stimulus. The uniqueness of these stimuli was limited by those CVC Word Lists contained in the Pilot Corpus. Only the meaningful words from the CVC Word Lists were used.

This test was conducted with a group of 11 subjects. These subjects were all members of ARCON's trained DRT listening crew. The test was repeated twice. The test was conducted over headphones in ARCON's test chamber. During the test, the CVC carrier sentences were presented, and each subject was asked to type the word that he/she perceived he/she heard into a data-entry unit. The presentation was electronically paced such that the listener crew could not proceed to the next sentence until all members had entered their response. The response set was totally open; the subjects were only told to enter the word they heard.

The results of this exercise are contained in the CVC Database as part of the Pilot Corpus.

14. Demonstration Material – Military Scenario

It was decided that a proper military scenario would allow for the generation of a meaningful demonstration of the ASE programs capabilities. The design requirements of this scenario included:

- Realistic situation
- Harsh acoustic environment to harsh acoustic environment
- Realistic military phrasing and vocabulary

Two U.S. Army officers, instructors at West Point who were stationed at MIT LL for the summer, developed an excellent scenario. The scenario involved communications between ground units and helicopter based units during the assault on a suspected enemy command center.

These officers participated in the initial exercise of this scenario. This was the first time that two MOTU units were used for data collection. An interface problem with these units caused timing problems across the various sensor channels. This was not discovered until after the officers had finished their summer assignment at MIT LL. The scenario was recorded with ARCON employees and is included as part of the Pilot Corpus.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
6 January 2005	Final Report	

4. TITLE AND SUBTITLE
Pilot Corpus for Multisensor Speech Processing

5a. CONTRACT NUMBER
F19628-00-C-0002

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
J. D. Tardelli

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420-9108

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency, Advanced Technology Office
3701 N. Fairfax Drive
Arlington, VA 22203-1714

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
ESC-TR-2004-084

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The estimation of speech parameters and the intelligibility of speech transmitted through low-rate coders are severely degraded when there are high levels of acoustic noise in the speaking environment. The application of nonacoustic and nontraditional sensors, which are less sensitive to acoustic noise than the standard microphone, is being investigated to address this problem, under the DARPA Advanced Speech Encoding program. Sensors investigated include the General Electromagnetic Motion Sensor (GEMS) and the Physiological Microphone (P-mic). In order to enable this research, a Pilot Corpus with simultaneous recordings from multiple sensors has been collected by ARCON Corporation, under subcontract to MIT Lincoln Laboratory. This report describes the corpus collection, including: corpus structure, acoustic noise environments, speech materials, the sensors, and baseline intelligibility evaluations. The corpus includes Diagnostic Rhyme Test (DRT) word lists, sentence lists, and Consonant Vowel Consonant (CVC) nonsense words. Noise environments include: M2 Bradley Fighting Vehicle (M2), Military Operations in Urban Terrain (MOUT), UH- 60 Blackhawk Helicopter (BH), and a Military Command Enclosure (MCE). This pilot corpus has been utilized by a number of DARPA-sponsored research teams for R&D on advanced speech encoding exploiting multiple sensors in the military noise environments.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
Unclassified	Unclassified	Unclassified	Same as Report	22	